BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 2: e25317
doi: 10.3897/biss.2.25317

OPEN ACCESS

Conference Abstract

Whip: Communicate and Test What to Expect from Data

Stijn Van Hoey‡, Peter Desmet‡

‡ Research Institute for Nature and Forest (INBO), Brussels, Belgium

Corresponding author: Peter Desmet (peter.desmet.work@gmail.com)

Received: 27 Mar 2018 | Published: 18 May 2018

Citation: Van Hoey S, Desmet P (2018) Whip: Communicate and Test What to Expect from Data. Biodiversity Information Science and Standards 2: e25317. https://doi.org/10.3897/biss.2.25317

Abstract

The ability to communicate and assess the quality and fitness for use of data is crucial to ensure maximum utility and re-use. Data consumers have certain requirements for the data they seek and need to be able to check if a data set conforms with these requirements. Data publishers aim to provide data with the highest possible quality and need to be able to identify potential errors that can be addressed with the available information at hand. The development and adoption of data publication guidelines is one approach to define and meet those requirements. However, the use of a guideline, the mapping decisions, and the requirements a dataset is expected to meet, are generally not communicated with the provided data. Moreover, these guidelines are typically intended for humans only.

In this talk, we will present 'whip': a proposed syntax for data specifications. With whip, one can define column-based constraints for tabular (tidy) data using a number of rules, e.g. how data is structured following Darwin Core, how a term uses controlled vocabulary values, or what the expected minimum and maximum values are. These rules are human- and machine-readable, which communicates the specifications, and allows to automatically validate those in pipelines for data publication and quality assessment, such as Kurator. Whip can be formatted as a (yaml) text file that can be provided with the published data, communicating the specifications a dataset is expected to meet. The scope of these specifications can be specific to a dataset, but can also be used to express expected data quality and fitness for use of a publisher, consumer or community, allowing bottom-up and top-down adoption. As such, these specifications are complementary to the core set of

© Van Hoey S, Desmet P. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

data quality tests as currently under development by the TDWG Biodiversity Data Quality Task 2 Group 2. Whip rules are currently generic, but more specific ones can be defined to address requirements for biodiversity information.

Keywords

data quality, fitness for use, requirements, specifications, documentation

Presenting author

Peter Desmet

Presented at

TDWG 2018 Annual Meeting